SIGNIFICANCE

Renal cell carcinoma (RCC) makes up over 90% of kidney cancers and currently is the most lethal genitourinary malignancy [1]. The incidence of RCC has nearly tripled in recent years in all races; however, the most dramatic increase is seen in African-Americans relative to other populations in the United States [1, 2]. According to the NCI Surveillance and Epidemiology End Results (SEER) Cancer Program [3], the age-adjusted incidence of kidney cancers compared to Caucasians is 30% greater (18.5 and 15.5 cases per 100,000 persons, respectively)(Figure 1). To date, research has not fully explained increased susceptibility to RCC among African-Americans [4]. Various hypotheses have been proposed implicating both genetic risk variants and a greater prevalence of RCC risk factors in African-Americans including obesity, chronic kidney disease, and hypertension [5, 6, 7, 8].

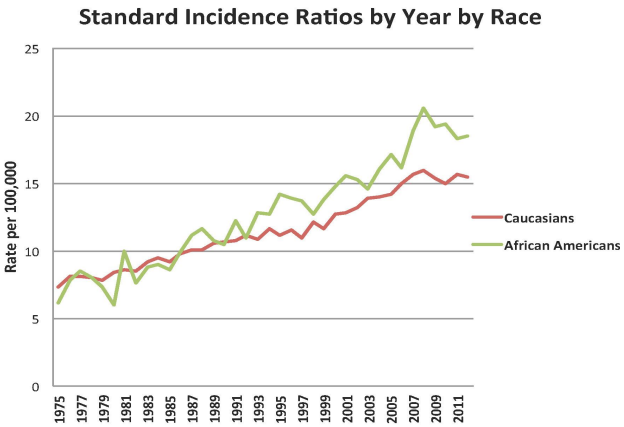


Figure 1: Standardized incidence ratios of cancer of the kidney and renal pelvis for Caucasians (Green) and African Americans (Red). Data from the Surveillance Epidemiology and End Result program from 1975-2011, [3])

Besides the higher incidence of RCC among African-Americans, several other racial disparities have been described. Some studies have demonstrated that African-Americans have a younger median age of RCC presentation, between 3-8 years earlier than Caucasians (Table 1). In RCC, age of onset is a major criterion for consideration for genetic testing as many hereditary cancers develop at a younger age than observed in the generation population [9]. While over a dozen known RCC syndromes exist, inherited risk and early disease onset may be more frequently related to a complex inheritance pattern. Specific risk alleles may contribute to the racial disparity in kidney cancer perhaps, predisposing to an earlier age of onset.

Series	Cohort	Black	Whites	*p value*
Kaiser Series	age ≥18	61.4 yrs	65.3 yrs	
		N=293	N=2152	<.0001
SEER Registry Cases 1988-2012	All Ages	61 yrs	64 yrs	
		N=18728	N=137927	<.0001
DOD Cohort (Lin et al.)*	age ≥18	53.1 yrs	61.2 yrs	
		N=370	N=2066	<.0001
CT State Registry	age ≥18	63 yrs	67 yrs	
		N=117	N=693	<.0001

**Calculated Median from Case Distribution*

Table 1: Age distribution of kidney cancer by race from a prior series [61] and ongoing work from the Yale Kidney Cancer group

An additional disparity in RCC is the large racial difference in the distribution of histologic subtypes. RCC is a group of cancers arising from the nephron with the two most common subtypes, clear cell RCC (ccRCC) and papillary RCC (pRCC) accounting for 85% of all cases. Although pRCC is considered to comprise 10-15% of renal tumors in general (Figure 2), several published and ongoing studies demonstrate this subtype is three-fold more common in African-Americans, accounting for 35-40% of cases (Table 2) [10, 11, 12]. The reason for the increased pRCC frequency in African-Americans is currently unknown. Unfortunately, when metastatic pRCC has an abysmal prognosis with limited therapeutic options.

Another major aspect of racial disparity in kidney cancer is that survival is also significantly worse among African-Americans. One explanation is that various studies have found that African-Americans less frequently receive standard treatments in the United States. However even controlling for treatment and tumor characteristics including, stage, grade, and subtype, survival is still significantly worse [4, 13, 14]. Similar to prostate cancer where African-Americans patients have a more aggressive disease biology,[15] it has

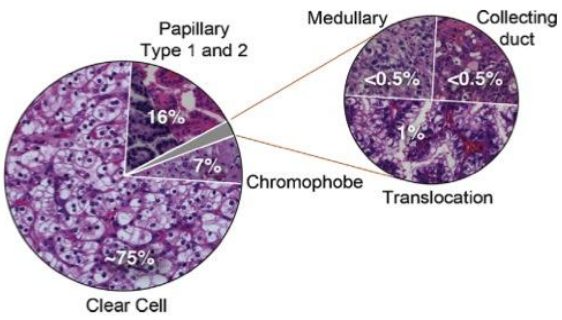


Figure 2: Histologic distribution of kidney cancer ([60])

been proposed that differences in molecular biology are involved racial disparities in kidney cancer [4, 6, 16, 17].

While significant racial differences exist in the incidence, mortality, age of onset, and subtype distribution of kidney cancer, no study has addressed genetic mechanisms associated with RCC racial disparity. While the Cancer Genome Atlas (TCGA) and other sequencing efforts have analyzed hundreds of kidney cancer specimens, the majority of tumors are from Caucasian patients. Without tumors from a diverse cohort of subjects it is difficult to explore to the reasons for these racial disparities including whether clinical differences are based on varying genomic backgrounds or specific driver alterations. With significant racial disparities present, we set out to identify possible coding and non-coding alterations explaining the genomic basis of kidney cancer racial disparity.

B. INNOVATION

In this work, we are interested in identifying key genomic alterations, which primarily contribute to the greater incidence, earlier age of onset, and different histologic distribution of kidney cancer in African-Americans compared to Caucasians. This study will be the first to comprehensively assess genomic alterations in kidney cancer by race. We expand upon prior work from the TCGA by including an additional cohort of African-American's with ccRCC. By including these samples and performing secondary data analysis of the existing ccRCC and pRCC datasets, we can compare differences in risk variants, driver mutations, and driver copy number alterations by race. Using our novel bioinformatic tools to analyze whole genome data, we will define and then validate non-coding driver alterations important in kidney cancer risk and progression. This study will be the initial step in addressing the biological/genetic causes of cancer health disparities in kidney cancer and the findings have implications far beyond the scope of this current proposal.

C. APPROACH

Aim 1: To perform whole genome sequencing (WGS) of African-Americans with ccRCC to complete a missing aspect of the cancer genome atlas (TCGA).

C-1-a Rationale: In recent years various TCGA efforts have characterized the genomic basis of the various forms of kidney cancer. These studies have led to the understanding that some of the diversity within kidney cancer results from different cells of origin giving rise to distinct types of cancer within the same organ. Additionally, differences in somatic alterations (driver mutations and copy number variations) are important in determining a cancer's molecular profile. In the TCGA, cases were submitted from various high volume tertiary centers to the Bio-specimen Core Resource (BCR) for accessioning and specimen processing. Specimens however were not submitted in a coordinated fashion to ensure the study population has a similar profile of that encountered nationally. Not surprising, there was clearly a limited number of African-Americans with clear-cell kidney cancer included in the TCGA analysis. Despite African-Americans accounting for approximately 1 in 7 cases of ccRCC, only a cursory analysis was performed in this population including 14/427 (3.3%) samples that underwent whole exome sequencing (Table 2) and 1/40 (2.5%) (Table 2) that underwent whole genome sequencing. Failing to include a larger population of African-Americans with clear cell RCC limits our ability to explore the genomic basis for racial disparities. With a higher incidence of pRCC in African-Americans, the papillary kidney cancer TCGA cohort was able to include a larger number of African-Americans. However, despite the available data, there has not been a thorough analysis of somatic driver alterations or germline risk variants more prevalent in African-American kidney cancer. We propose to complete the TCGA analysis of the top two subtypes of kidney

EXOME SEQUENCING DATA					
		Total	Black	White	Other/NA
TCGA Clear Cell RCC	#	427	***14***	400	13
	%	100%	***3.3%***	93.7%	3.0%
TCGA Papillary RCC	#	159	42	100	17
	%	100%	26.4%	62.9%	10.7%

WHOLE GENOME SEQUENCING DATA					
		Total	Black	White	Other/NA
TCGA Clear Cell RCC	#	40	***1***	36	3
	%	100%	***2.5%***	90.0%	7.5%
TCGA Papillary RCC	#	32	14	13	5
	%	100%	43.8%	40.6%	15.6%

Table 2: Racial and histologic distribution of available whole exome and whole genome data available from TCGA datasets

cancer, papillary and clear cell, by analyzing an additional cohort of African-Americans with ccRCC. By performing whole genome sequencing on this additional cohort of samples, we will have an adequate number of cases to allow balenced comparisons between African-American and Caucasian clear cell and papillary kidney cancers.

C-1-b Sample acquisition and DNA extraction: All patients undergoing scheduled kidney cancer surgery at Yale New Haven Hospital are offered enrollment into an IRB-approved Genitourinary Biospecimen repository (P.I. Shuch, HIC# 0805003787). Within 30 minutes of removal, fresh tumor tissue is snap frozen in liquid nitrogen by the pathology team. Additionally whole blood is procured to serve as a genomic control. In the past 2 years, over 300 subjects with kidney cancer have been prospectively enrolled. All fresh bio-specimens are stored at -80˚C and are available for immediate analysis. For the purpose of completion of the TCGA dataset, we will utilize a consecutive series of 15 African-American subjects with ccRCC from 2013-2015. DNA will be extracted from fresh tumor tissue and whole blood using an automated Maxwell 16® System (Promega, Madison, WI).

C-1-c WGS and variant calling: Sequencing of the normal and tumor sample will performed using Illumina's Hiseq 2000 technology. In brief, DNA fragments from each sample will be hybridized using HiSeq Paired-End cluster Kits and will be further amplified using the Illumina cBOT. Paired–end libraries will be generated by utilizing HiSeq (2x101) cycle and imaging will be performed by TruSeq kits.

We have extensive experience in large-scale variant calling and interpretation through being active members of the 1000 Genomes Consortium, especially in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups of the consortium where the majority of the variant calling tools were developed, deployed and interpreted [18, 19, 20]. We have already developed a prototype pipeline for calling germline and somatic variants. We will use the Genome Analysis Toolkit (GATK)[21] to call germline SNPs and INDELS. We use parmaters consistent with those used in TCGA[22].We will map raw FASTQ files of each sample to the hg19 reference genome using bwa-mem algorithm with default parameters to generate BAM files. These bam files will be further processed to sort and mark duplicates reads before calling variants.

We will follow GATK best practices [21] to generate initial raw variant call sets using GATK haplotype caller. We will filter these initial call sets by running GATK variant recalibration tool. The filtering strategy based on variant recalibration method uses a continuous adaptive error model. The adaptive error model takes into account the relationship between annotation of each variant (Quality score, mapping quality, strandedness and allele information) and the probability of it being a true positive instead of a sequencing artifact. Furthermore, we will exclude any filtered variant, which falls in a low mappability region of the genome. In addition, we will utilize MuTect [23] and Strelka [24] to call somatic SNVs and INDELs, respectively.

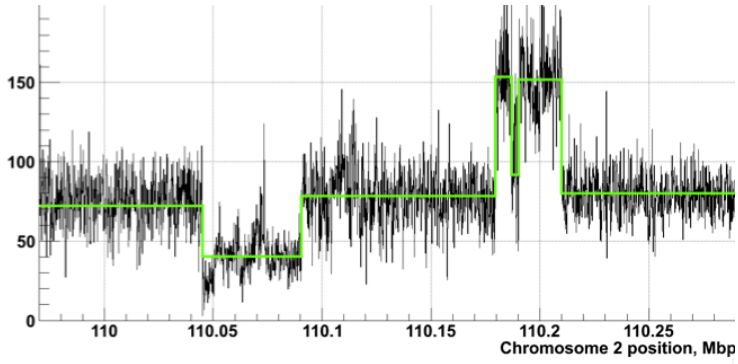


Figure 3: Read depth based identification of copy number variation bv CNVnator.

Structural variations (SVs) are important contributors to human polymorphism, have great functional impact and are often implicated in various diseases including cancer. We have developed a number of SV calling algorithms, including BreakSeq [25], which compares raw reads with a breakpoint library (junction mapping) , CNVnator, which measures read depth[26], AGE, which refines local alignment [27], and PEMer, which uses paired ends [28]. We have also developed array-based approaches [29] and a sequencing-based Bayesian model [30]. Furthermore, we have intensively studied the distinct features of SVs originated from different mechanisms. This indicates specific creation processes and potentially divergent functional impacts [31, 32]. We will perform extensive molecular characterization of germline and somatic SVs in these cancer samples. We will run CNVnator to identify germline and somatic copy number variations in each cancer samples. We will apply CREST [33] to generate germline and somatic large structural variations including large deletions, insertion, inversion, intra & inter-chromosomal translocations. Furthermore, we will run our BreakSeq tool to decipher the underlying mechanism of somatic and germline SV formation.

C-1-d Deliverables: In this aim, we will generate an extensive catalogue of germline and somatic variants including SNPs, INDELs and large SVs for African-American ccRCC cases. This will be done consistently with the methodology already used in the TCGA, so this catalogue can be used conjunction with TCGA kidney cancer genomic variant datasets to serve as an excellent comparison for the identification of genomic aberrations, associated with racial disparity observed in the emergence of kidney cancer. We plan to make our sequencing data available via dbGAP (see data dissemination plan).

Aim 2: To assemble a set of coding and non-coding regions associated with kidney cancer, both in terms of somatic and germline alterations.

C-2-a Rationale: In this study, we aim to discover underlying genetic regions that explain racial disparity in RCC. However, due to the limited size of sequenced samples, it is not feasible to test every single region in the genome. In fact, as we discuss later, we have to limit our search space to achieve sufficient statistical power. Therefore, we will first assemble a catalog of mutations that are relevant to renal cell carcinoma and prioritize regions with greatest impact. In this way, we will incorporate our best prior knowledge of RCC and cancer genomics into this study. This allows us to decrease the number of tests, avoiding losing statistical power.

C-2-b Relevant Preliminary Results

C-2-b-1 We have developed ways of prioritizing high-functional impact variants: We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets [34, 35, 36]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations [34]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region[37]. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery) [38, 39]. In recent studies [31, 40], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq. It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). It identifies deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their specific disruptiveness to TF binding sites, annotating both loss-of and gain-of function events. Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast, and prostate cancer samples [31]. We have also applied our method to investigate non-coding mutation patterns in subtypes of gastric cancer [41]. Drawing on this experience, we are currently co-leading the International Cancer Genomics Consortium (ICGC) pan-cancer analysis-working group (PCAWG)-2 (analysis of mutations in regulatory regions) group.

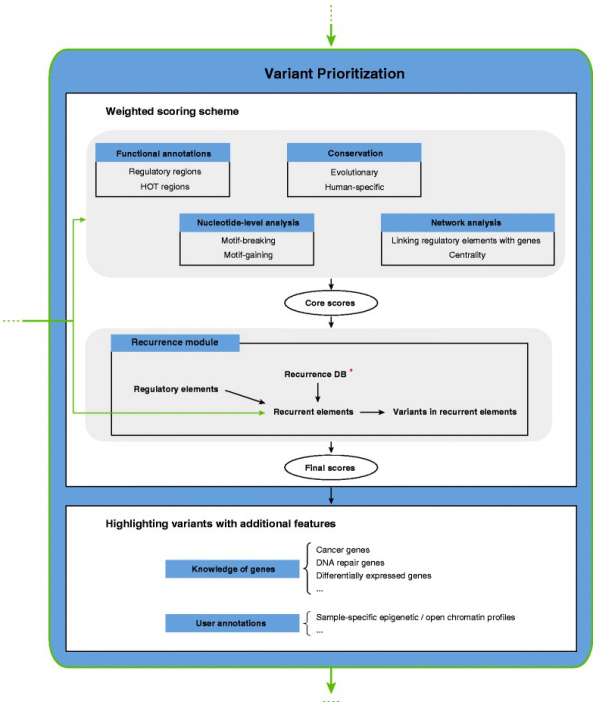


Figure 4: Workflow for Funseq based variant prioritization

We have also used allelic variability to prioritize regions of the genome. That is we prioritize regions that differ in functional genomic response, for example, allele-specific expression and binding, between the maternal and paternal alleles. Our variant analysis work includes AlleleSeq [42], a computational pipeline to identify allele-specific events, and AlleleDB, our database connecting single nucleotide variants with allele-specific binding and expression [43].

C-2-b-2 We have developed tools for somatic and germline burden tests: We have worked on statistical methods for analysis of non-coding regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in noncoding elements by comparing observed mutation counts with expected counts under a whole genome background mutation model. LARVA includes corrections for biases in mutation rate owing to DNA replication timing. LARVA can also be used in a mode exclusively on coding regions to prioritize genes. We used this tool in a pan-cancer analysis of 760 cancer whole genomes' variants spanning a number of cancer data portals and some published datasets. Our analyses demonstrated that LARVA can recapitulate previously established coding and noncoding cancer drivers, including the TERT and TP53 promoters [44].

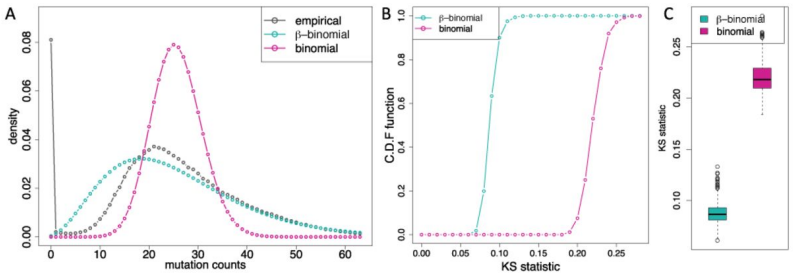


Figure 5: Comparison of β-binomial distribution fit (turquoise) and binomial distribution fit (pink) to observed cancer somatic mutation counts. The β-binomial distribution betters models the empirical distirbution's (black) overdispersion.

C-2-b-3 We have already identified some regions associated with kidney cancer through our involvement in the papillary TCGA team: Due to Yale's expertise in the clinical management and genetics of kidney cancer, we were invited to participate in the various TCGA kidney cancer projects. Our role in the TCGA KICH (chromophobe RCC) included being the manuscript coordinator for the Cancer Cell manuscript. Next for the TCGA KIRP (pRCC) our team performed the analysis of the whole genome sequencing, now published in New England Journal of Medicine [22], providing us with further experience with the available RCC genomic datasets. Finally our team has participated in two ongoing pan-RCC manuscripts serving a central role assessing evaluating the cluster of clusters' (Cluster of cluster assignments -- COCA) immunologic profile from gene and miRNA expression datasets. Together with other published results on RCC [45, 46, 47, 48, 49], we already have the ability assembled an extensive list of impactful regions on the genome that have shown statistical significance in previous studies. However, most of these studies focused on coding regions only.

C-2-c Research plan

C-2-c-1 Collected database of what is known for kidney cancer: First, we will mine the literature and condense results from previous studies, alluded to above. We will gather genetic changes that include but not limited to: single nucleotide variation (SNP), structural variation/copy number variation (SV/CNV), and mutation process signature. Our study will take a comprehensive approach of the entire genome, sweeping a larger pool to unearth genomic regions for racial disparity in RCC. Last, we will also notate known regions that have discovered in studies on Caucasian and African-American disparity in other types of cancer, for example, prostate cancer [50, 51]. That is, we will add to our collection of kidney-cancer prioritized regions, regions known to be associated with racial disparities in other types of cancer. Studies have pointed out that RCC is uniquely characterized by copy number variations (RCC) as early and major driver event [45]. Repeats are triggering factors for many structural variation events. Therefore in the germline analysis below, we will also pay attention to repeats polymorphisms around pre-eminent cancer associated genes and recurrent CNV regions in RCC. Particularly, we assume excessive repeats put certain RCC related genes predisposed to CNV events.

C-2-c-2 We will extend FunSeq to find connected modules of elements: FunSeq has already had a limited way of connecting non-coding elements to target genes. It exploits locality of promoters and correlates epigenetic markers on distal regulatory elements with gene expression. Here will extend this capability to develop modules of elements. Genetic modules extend high impact regions by linking them with other genomic elements according to physical interaction, epigenetic marker and expression correlation, molecule pathway/network and other evidence. Elements in the same genetic module are expected to play similar roles

in ccRCC and pRCC initialization and development. In the end, we will integrate this new feature into FunSeq and use to assemble genetic modules. Genetic modules group potentially impactful elements that share similar or collaborative biological functions, increasing statistical power in our study. Last, genetic modules offer annotation to less known noncoding regions. Our results will be more biologically interpretable because these regions will be linked with genes.

In order to systematically integrate evidence from various sources (which can mostly be represented in graph form), we will use a random walk on multiple graph layers. At each step, we chose to update the state on one graph. The walk stops at certain distance from its starting point (boundary condition). Starting from the gene that we are interested with and simulating this random walk multiple times, we will finally tally the number of visits to each node and pick out "hot" nodes (that are often covered in walks). Those nodes represent the linked nodes with our starting gene. Since random walk will give an empirical distribution of number of visits to the nodes, we will be able to set up our cut-off value for linked nodes in a robust manner.

C-2-c-3 We will extend LARVA to include additional covariates: It is known that various genomic features affect background mutation rate in most cancer types, which results in numerous false positives in somatic mutation recurrence analysis [50]. Hence, we have been working on an update to LARVA, which incorporates corrections for additional covariates that influence the somatic mutation rate in different genomic regions, including sequence content, replication timing, expression level, histone modification marks, and chromatin status. Our intention is to iteratively refine the underlying whole genome background mutation model to reflect all factors that influence the accumulation rate of background mutations.

C-2-c-4 We will run our updated and extended FunSeq & LARVA on WGS sequences from TCGA and aim 1: We expect many changes in noncoding regions play a critical role in renal cell cancer. In order to find high impact mutations in noncoding regions, we will run our updated and extended FunSeq and LARVA on variation calls from TCGA whole genome sequenced samples as well as our newly sequenced samples, both cancer and normal. In a first pass run, we have already run FunSeq on 32 whole genome sequenced samples from the TCGA KIRP group. We have found several disruptive mutation hot spots in the genome. We also have found excessive somatic mutations in MET intronic and promoter regions, along with several other recurrent mutated regions that merit further investigation.

C-2-c-5 We will identify critical regions burdened by germline mutations: First, we will try to find regions that are significantly burdened by germline mutations in the kidney cancer cases versus healthy people. As a non-cancer control, we will use both the 1000 Genome Project (2504 people) for the whole genome as controls as well as the Exome Aggregation Consortium (ExAC, non cancer) for the exome[51]. We will look for regions and genes that are burdened significantly in RCC, compared to the control. Given the size of the datasets, we will be well powered in our tests (comparing numbers discussed later in aims 3 and 4). We will also prioritize regions that are less significantly associated with RCC that are known to have racial disparities in healthy people.

C-2-d Deliverables: We aim to generate a list of regions in genome that, to our best knowledge, potentially have the highest impacts on the development of RCC. We will also construct a list of genetic modules that are assembled from high impact regions. We will make these regions available from our project website and as tables in published papers (see data dissemination plan). In Aim 3, we will directly test those elements on our samples.

Aim 3: To identify genomic alterations differing most between African-Americans and Caucasians with kidney cancer

C-3-a Rationale: In this aim, we are going to take the genomic regions and modules that we have developed in Aim 2 and test for any evidence of racial disparity. Our overall intention is to investigate differences in the occurrence of common SNPs or differential burdening in terms of rare germline or somatic mutations between African-American and Caucasians. Our specific goal for this aim is to score and prioritize these genomic regions in order to select 550 regions with the highest score to be validated in Aim 4 using a larger cohort.

C-3-b Compare the germline mutations in coding regions between Caucasian and African-Americans in the prioritized regions using WES Data

C-3-b-1 Variant level analysis: For the coding region analysis, we will utilize the full category of 556 samples with whole exome data analysis from TCGA. For the common variants analysis at a single locus, Fisher's exact test can be used to evaluate the racial disparity between Caucasian and African-American subjects with RCC. Here, we prioritize common variants according to their associations with RCC disparity in race. For a common SNP identified in African Americans and Caucasians with RCC, we record the minor allele frequencies and major allele frequencies in African Americans and Caucasians with RCC, respectively. For these counts of the focal SNP, Fisher-exact test is used to determine whether the SNP tends to be associated with the African Americans with RCC. The p-values of the tests for all the common variants are used to prioritize them for further study and validation. Moreover, the power of the Fisher exact test can readily be estimated in this context. For instance, for an ordinary SNP with allele frequency 7% in the total samples, when its frequency in the African American subjects is 12%, the power of the test can reach 0.4 with a p-value < 5e-5. This indicates that these SNPs can be detected with statistical significance from 1000 candidates, even when a Bonferroni correction is used.

C-3-b-2 Region based analysis: Beyond investigating the association between the single common variant and race, we will focus on the evaluating the cumulative effects of a set of rare variants in certain genomic regions, such as genes, using both burden and non burden test. Burden test are often applied on regions where most of the variants in the same region are causal and their effect on the phenotype are on the same direction. We assume that in total there are n patients with whole exome sequencing data available. Also for a target region, for example, a gene, there are m variants. Let y_i denote the population information of the i^{th} patient. $y_i = 1$ for African-Americans and 0 otherwise. Let $\mathbf{G_i} = (g_{i1},...,g_{im})'$ represent the genotype of patient i. Then a logistic regression model can be set up to evaluate the association as in (1). Suppose that π_i describes the mean of the population status, then

$$\text{logit}(\pi_i) = \gamma_0 + \mathbf{G_i'b} \qquad (1).$$

For the burden test, we could treat the coefficient b_j for each patient as a weighted coefficient like $b_j = w_j \times b_c$. Then equation (1) can be rewritten to

$$\text{logit}(\pi_i) = \gamma_0 + b_c \left\{ \sum_{j=1}^{m} w_j g_{ij} \right\} \quad (2).$$

Then under the null hypothesis that there is no association of variants in this region with race, the coefficient b_c should be zero. So the test statistic for H_0: $b_c = 0$ should be

$$Q_B = \left[\sum_{i=1}^{n} (y_i - \hat{\pi}_i) \left(\sum_{j=1}^{m} w_j g_{ij} \right) \right]^2 \qquad (3).$$

The allele frequency can be used to assign the weight for each variant. For example, $w_j = 1 \Big/ \sqrt{\hat{p}_j (1 - \hat{p}_j)}$, where $\hat{p}_j$ is the minor allele frequency. However, in some cases, where the target region has many non-causal variants or the effect of such variants is quite heterogenous, burden tests, such as equation (3), may lose statistical power. Here, sequence kernel association test (SKAT) can be used. Instead of assuming a weighted coefficient effect in the burden test, b_j s are treated as independent random variables with 0 mean and variance $w_j^2 \tau$. Then the null hypothesis can be changed to H0: $\tau = 0$. Then the test statistic under equation (1) can be written into:

$$Q_S = (\mathbf{y} - \boldsymbol{\pi})' \mathbf{K} (\mathbf{y} - \boldsymbol{\pi}) \qquad (4).$$

In (4), $\mathbf{K} = \mathbf{GWWG}'$ is the kernal matrix, and $\mathbf{G}$ is the genotype information vector. $\mathbf{W} = diag\{w_1, ..., w_m\}$ is the weight matrix which can choose based on allele frequency or external information, such as conservation score. The test statistic in (4) can be rewritten into

$$Q_S = \sum_{j=1}^{m} w_j^2 S_j^2 = \sum_{j=1}^{m} w_j^2 \left\{ \sum_{i=1}^{n} g_{ij} (y_i - \hat{\pi}_i) \right\}^2 \quad (5).$$

In the coding variant analysis, because for most genes, we do not know which of the two cases each gene falls into, a unified test can be used as the following

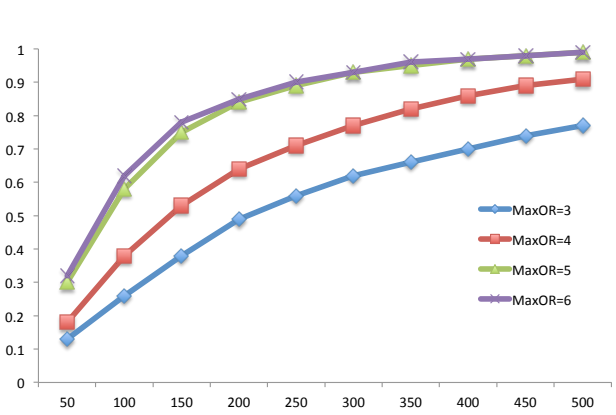
$$Q_\rho = \rho Q_B + (1-\rho) Q_S, 0 \le \rho < 1 \qquad (6)$$

Since the best route in (6) is unknown, the best test statistic can be used as the following:

$$Q_{opt} = \min\left(Q_{p1}, \cdots, Q_{p_k}\right) \qquad (7)$$

C-3-b-3 Power analysis using SKAT for per region based analysis: In the above, we are planning to use aggregated burden tests (i.e. SKAT) to look for differential burdening between populations and use this to rank the regions. While we are not striving for absolute statistical significance in the differential burdening, we do think that the sample size is enough to get an appreciable signal for ranking. Here, we discuss the power aspects of the burden tests in detail and substantiate. SKAT analysis has been developed for rare genomic mutations and remains robust even for common variants. We will utilize SKAT to identify genomic regions with significant variant disparity in kidney cancer between Caucasian and African-American populations. To estimate the sample size that we need to use in order to obtain statistical power, we used the SKAT package from R project, running several population models with different parameters (Figure 6). In the proposed study, we will focus on genomic modules linked with kidney cancer; therefore we expect a greater number of effective mutations.

A



B

Max OR	3	4	5	6
effective sample size	763	326	172	160

Figure 6: Using the default haplotype information in the SKAT.haplotypes dataset, we randomly selected subregions of size=5k and ran 100 simulations. In A, we show the statistical power obtained across the different models of maximum Odds Ratio. In B) we show the required sample size for each of these models in order to obtain significant statistical power (α=0.01, β=0.2)

C-3-c Compare the germline mutations in noncoding regions between Caucasian and African-Americans in the prioritized regions using WGS Data

C-3-c-1 Pooled variant test for limited target regions: For the noncoding region analysis, since we only have limited power with 32 WGS samples in both populations, target regions instead of the whole genome wide analysis will be carried out on only a small set of regions. From our experience with TCGA KIRP, we already prioritized MET intronic and promoter regions, along with several other recurrent mutated regions that merit further investigation. We will only focus on these selected regions to use the unified test mentioned above (in section C-3-b).

C-3-c-2 Non-parametric test for FunSeq score distribution difference: We suspect that the casual regions may not only be under differential mutational burden between races, but may also be overly affected by high-impact mutations. Thus, for the prioritized regions given above, we plan to calculate all the FunSeq scores on both African-American and Caucasian populations. Subsequently, by ranking and pairing the scores between the two population groups we intend to use the Wilcoxon signed-rank test to evaluate the significance of the mutational impact on each region. This test is a non-parametric version of the paired t-test and is used when we cannot assume that the populations follow a normal distribution. As the population size increases, a Z-score can then be calculated.

C-3-d Compare the somatic mutations between Caucasian and African-Americans in the prioritized regions: Previously, we developed an integrative framework LARVA to discover the highly recurrent regions in cancer genomes as candidates for drivers [43]. We will further develop our method by correcting many other genomic features for more accurate background mutation rate calculation. Specifically, in a region with length l, suppose the mutation rate is known as π, then the number of mutations y within l given μ should follow a Poisson distribution as the following:

$$p_Y\left(y|\mu\right)=\frac{e^{-\mu}\mu^{y}}{y!} \qquad (8)$$

However, we discovered in our previous analysis that there is great cancer type, sample, and regional heterogeneity in the mutation count data [43]. Such mutational heterogeneity violates the constant mutation rate assumption and results in over-dispersion. Hence, instead of supposing μ is constant, we set up the following model

$$\begin{aligned} p_Y\left(y|\mu\gamma\right) &= Poisson\left(\mu\gamma\right) \\ \gamma &\sim Gamma\left(1,\,1/\sigma^{2}\right) \end{aligned} \qquad (9)$$

Then the marginal distribution of Y could be expressed as the type I negative binomial distribution

$$p_Y\left(y|\mu,\sigma\right)=\frac{\Gamma\left(y+1/\sigma\right)}{\Gamma\left(1/\sigma\right)\Gamma\left(y+1\right)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{y}\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma} \qquad (10),$$

Where $E\left(Y\right)=\mu$, $Var\left(Y\right)=\left(1+\sigma\right)\mu$. Let $x_1,x_2,\cdots,x_k$, are the genomic covariates to be corrected, such as replication timing, GC content, and chromatin status, we could use the following negative binomial regression to estimate the local mutation rate under the covariant set.

$$\begin{aligned} g_1\left(\mu\right) &= \log\left(\mu\right)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k \\ g_2\left(\sigma\right) &= \log\left(\sigma\right)=\alpha_0 \end{aligned} \qquad (11)$$

Consequently, instead of estimating a genome wide mutation rate m we are now estimating a coefficient vector for the mean and a constant over-dispersion value. For each region to be estimated, a local mutation rate can be reconstructed by equation (11) for accurate background rate and false positive/negative controls.

We will apply our new method on the 16 African-American and then the 16 Caucasian WGS samples separately. Highly recurrent regions in each will be reported and compared. Those regions that are unique to either population will be prioritized for detailed validation.

C-3-e Deliverables: This aim will create a ranking on the list of genes, non-coding regions and variants from Aim 2 to pass to the validation in Aim 4. We will combine the rankings from the sections above by comparing

their corresponding p-values. However, we will keep a minimum number of validations of each category. Also, we plan to make our racial disparity rankings of genes and non-coding regions publicly available from the project web server (see data dissemination plan).

Aim 4: To validate specific regions with either germline or somatic mutations suspected of contributing to kidney cancer racial disparity.

C-4-a Rationale: Typically, traditional GWAS studies require thousands of sequenced genomes to associate genetic variants and disease with confidence. Therefore, aim 2 and 3 may not render the necessary statistical support to associate kidney cancer variants with racial disparities. However, in aim II and III, our main intention is to obtain, prioritize and rank ~550 genomic regions using our Funseq algorithm [40]. These regions will be further processed and validated in aim 4, by using sequences from Yale's Genitourinary Biospecimen Repository. We intend to validate 55 regions (100bp each) for 384 individuals.

In particular, we will assemble a independent validation cohort from Yale's Repository. This will contain both African-Americans and Caucasian with clear cell and papillary RCC to allow comparisons across both histologic type and race. Besides confirming an association with kidney cancer, a large cohort will help us better understanding of how frequently these alterations occur.

C-4-b Power analysis for the validation cohort: Here we focus on 550 common SNPs prioritized by the Fisher exact test proposed in Aim 3. The same test is adopted to detect the SNPs associated with racial disparity of RCC, using the equal number (192) of African American and Caucasian patients with RCC. To analyze the test power, we survey the parameter space of a candidate SNP, i.e. the frequencies of the SNP in total patients (f) and in the African American (f_a) and Caucasian (f_c) patients. Due to correcting multiple tests with Bonferroni method, only SNPs with p-value < 1.0e-4 are considered to be associated with race disparity of RCC. Using STATMOD package [52], we find that to be detected with power at 0.8, a candidate SNP requires its f and f_a /f_c larger than 0.08 and 3.5 respectively. However, note, Bonferroni correction is overly stringent, rendering this power analysis conservative.

In the other extreme, when all the prioritized regions are genes after pooled rare SNP tests, we suppose eventually 10 genes with 5kb length. Using the SKAT R package, we performed a power analysis of 100 simulations and we could still detect regions with an Odds Ratio (OR) equal to 4 with this number of samples (power > 0.8).

C-4-c Sample acquisition and DNA extraction: As mentioned above, fresh kidney cancer tissue is procured on our IRB-approved Genitourinary Bio-specimen protocol within 30 minutes of removal. Additionally our protocol allows access to archival tumor tissue from 1988-2013. Yale pathology archives have available formalin-fixed, paraffin-embedded (FFPE) tissue blocks to retrieve tumor and the adjacent normal kidney tissue for a genomic control. All tumors have recently been centrally reviewed by our genitourinary pathologists and classified according to recent International Society of Urologic Pathology (ISUP) criteria [53]. For our validation cohort, an equal number (n=96) of Caucasian and African-American clear cell and papillary tumors (total n=384) will be selected as a Yale Validation Cohort. For both fresh and FFPE tissue, DNA will be extracted using an automated Maxwell 16® System (Promega, Madison, WI).

C-4-d Genotyping kidney cancer non-coding genomic variants: Frequently, next generation sequencing results find variants that require validation to confirm significance. We will employ similar methods to various other studies involving novel variants found on either exome or whole genome sequencing [54, 55, 56, 57] These studies utilize the MassArray system (Agena Biosciences, San Diego, CA), a mass spectrometry platform that measures PCR-derived amplicons. The system can be multiplexed to analyze a large number of alterations with high sensitivity but a low cost. DNA from FFPE and fresh tissue can be analyzed on the same chip without difficulty. The MassArray system has been utilized to study genomes in the study of cancers, benign conditions, and even ecologic research. This approach has been used in the screening of large case-

control series and independent validation cohorts of affected individuals. For the analysis, 20 ng of DNA from genomic control will be used in Yale's Validation cohort. The 384-subject cohort will have their germline DNA assessed for non-coding variants identified. The MassArray Assay Design Suite will be used for designing custom PCR primers to detect potential germline variants using the genomic coordinates of interest. For genotyping, up to 40 genomic variants can be multiplexed per well per silicon chip. Mutation calls for each tumor and germline sample will be assessed using the MassArray Typer 3.4 Analyzer. Unlike next generation sequencing, the results can be quickly automated and generated into both a graphical or table interactive format.

C-4-e Tumor profiling for somatic non-coding mutations: The MassArray system is frequently used for the rapid detection of known or suspected somatic alterations important in cancer [58, 59]. Panels exist to detect common alterations in specific cancers and are employed for testing at various clinical laboratories. We will perform somatic mutational profiling using the MassArray System that allows multiplexing for up to 15 somatic variants per 384-well chip. The technology can detect variants with as low as 1% mutant allele frequency using a small DNA quantity. A total of 20 ng of total DNA will be obtained from tumor DNA from the Yale Validation Cohort to determine if racial differences exist in somatic non-coding mutations between African-Americans and Caucasians with kidney cancer. Small insertions/deletions or single nucleotide alterations found from the WGS and secondary TCGA data analysis will be assessed in the validation cohort. For somatic variants, the MassArray Assay Design Suite will be used for designing custom PCR primers. Similar to above, mutation calls will be assessed using the MassArray Typer 3.4 Analyzer.

Figure 7: Workflow for validation of whole genome sequencing findings using Yale cohort of tissue. DNA from formalin-fixed, paraffin embedded tumor tissue and also genomic controls will be amplified, have primer extension, mass spectroscopy detection, and analysis. Germline and somatic, coding and non-coding variants will be validated with a large Yale patient cohort.

D. Potential Pitfalls and Alternative Strategies

One potential issue we see with our proposal is that, despite our design, we may not find any regions in the validation cohort that are significantly different between races. We have designed the study so that we believe we will have adequate power to detect racial differences but, of course, we will not know until we get to the validation. If after doing the first third of the validations (~185) we are not finding any regions that are significantly different, we have a number of courses of action: (1) We can remove somatic variants from the validation. Validating somatic variants is more expensive than germline ones. Removing them will allow us to validate a potentially larger number of regions. (2) We can focus only on disparities in coding genes as opposed to non-coding regions. There are many more kidney cancer exome sequences done than WGS (by more than an order of magnitude) and coupling this with the much smaller genomic space being queried should substantially increase the power of our analysis. (3) The validation cohort can be expanded to increase power. Currently the Yale Biospecimen Repository is adding 150 new kidney cancer subjects each year. Additionally our close collaborator in the US Kidney Cancer Study has access to a large cohort of genomic DNA in individuals with kidney cancer (843 Caucasians and 358 African-Americans). Finally if needed, the Yale Kidney Cancer Program recently was granted approval from the Connecticut State Tumor Registry to access records and/or tissue from individuals with a diagnosed with kidney cancer from 1998 to present.

E. Expected Outcomes and Future Directions

At the conclusion of this research, our analyses will determine candidate coding and non-coding regions associated with papillary and clear cell kidney cancer. We will identify and then validate specific germline and somatic alterations that are disparate in their distribution in African-Americans and Caucasians with kidney cancer. These findings will be an initial step towards understanding the genomic cause of kidney cancer racial disparity and have implications beyond the scope of this project. Understanding inherited predisposition to kidney cancer may have important screening implications in high-risk individuals such as African-Americans. Additionally, racial disparity in candidate driver alterations has the potential to impact how we view treatment in the age of precision-based cancer therapeutics. The findings from this project have far reaching implications, justifying further research beyond the scope of this proposal.